Neurocomputing 73 (2010) 2332–2345

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A modified Markov clustering approach to unsupervised classification of protein sequences

László Szilágyi [a,b,*], Lehel Medvés [a], Sándor M. Szilágyi [a]

[a] Sapientia University of Transylvania, Faculty of Technical and Human Sciences, Şoseaua Sighişoarei 1/C, 547367 Corunca, Romania
[b] Budapest University of Technology and Economics, Department of Control Engineering and Information Technology, Magyar tudósok krt. 2, H-1117 Budapest, Hungary

ARTICLE INFO

Available online 2 June 2010

Keywords:
Markov clustering
Bioinformatics
Protein sequence classification
Unsupervised classification
Sparse matrix

ABSTRACT

In this paper we propose a modified Markov clustering algorithm for efficient and accurate clustering of large protein sequence databases, based on previously evaluated sequence similarity criteria. The proposed modification consists in an exponentially decreasing inflation rate, which aims at helping the quick creation of the hard structure of clusters by using a strong inflation in the beginning, and at producing fine partitions with a weaker inflation thereafter. The algorithm, which was tested and validated using the whole SCOP95 database, or randomly selected 10–50% sections, generally converges within 12–14 iteration cycles and provides clusters of high quality. Furthermore, a novel generalized formula for the inflation operation is given, and an efficient matrix symmetrization technique is presented, in order to improve the partition quality with relatively low amount of extra computations. Finally, an extra speedup is achieved via excluding isolated proteins from further processing. The proposed method performs better than previous solutions, from the point of view of partition quality, and computational load as well.

1. Introduction

One of the main goals of functional genomics is to establish protein families in large databases. Successful classification of protein families can have significant contributions to the delineation of functional diversity of homologous proteins, and can provide valuable evolutionary insights as well [9].

By definition, protein families represent groups of molecules showing relevant sequence similarity [4]. Members of such protein families may serve similar or identical biological purposes [13]. Identifying these families is generally performed by clustering algorithms, which are supported by similarity and/or dissimilarity measures, previously computed between all pairs of protein sequences.

Performing an accurate clustering results in such protein families, whose members are related by a common evolutionary history [10]. If this condition holds, well established properties of some proteins in the family may be reliably transferred to other members whose functions are not well known [12]. Several protein clustering methods are currently available in the literature [7,21]. One of the greatest obstacle for them represents the multi-domain structures of many protein families [5].

TRIBE-MCL is an efficient protein sequence clustering method proposed by Enright et al. [9], based on Markov chain theory [6,7]. The authors assigned a graph structure to the protein database such a way that each protein has a corresponding node, while initial edge weights in the graph represent a priori computed similarity values, obtained via BLAST search methods [1]. Clusters were then obtained by alternately applying two operations to the similarity matrix: inflation and expansion. The former represents a task, which reevaluates the values within columns of the matrix by raising higher probabilities and suppressing the small ones, while the latter aims at favoring longer walks along the graph, which is obtained via matrix squaring.

In a previous version of this paper [15] we proposed a modification of the TRIBE-MCL algorithm, in order to enhance its accuracy, improve its reliability and weaken its computational load. These were achieved by introducing a time varying inflation rate and thus forcing the algorithm to apply a stronger inflation in the first iterations when the hard structure of the clusters is established, and reducing inflation strength for fine tuning the cluster shapes in later iterations. In this paper we introduce a singleton filter, which improves the execution speed of the algorithm without influencing the outcome of the clusters. Further on, we propose and apply a complete methodology for the qualitative and quantitative evaluation of the final results.

* Corresponding author at: Sapientia University of Transylvania, Faculty of Technical and Human Sciences, Şoseaua Sighişoarei 1/C, 547367 Corunca, Romania. Tel.: +40 265 208170; fax: +40 265 206211.
E-mail addresses: lalo@ms.sapientia.ro (L. Szilágyi), medveslehel@yahoo.com (L. Medvés), szs@ms.sapientia.ro (S.M. Szilágyi).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.02.023

The remainder of this paper is structured as follows: Section 2 takes into account the functional details of the TRIBE-MCL algorithm and presents our motivations to introduce modifications. Section 3 presents the details of the modified TRIBE-MCL algorithm. Section 4 evaluates and discusses the efficiency and accuracy of the proposed method, making an exhaustive comparison with several constant inflation cases. Section 5 presents the conclusions and gives some hints for further research.

2. Background

2.1. The SCOP database

The structural classification of proteins (SCOP) database [19] was developed as an evolutionary classification, focusing on the creation of a coherent evolutionary framework of proteins, based on their conserved structural features. The hierarchical classification within the SCOP database is based on protein domains, which are defined as evolutionary units observed in nature either in isolation or in more than one context in multi-domain proteins [2]. The domain hierarchy is structured into levels of families, superfamilies, folds and classes. The reader interested in such details is referred to [16,3,14].

The SCOP95 database that is involved in this investigation, is a subset of SCOP (version 1.69), which contains 11 944 proteins, exhibiting a maximum similarity of 95% among each other. Several precomputed similarity and distance matrices (BLAST [1], Smith–Waterman [20], Needleman–Wunsch [17], PRIDE [11], etc.) are available at the protein classification benchmark collection [18].

2.2. The TRIBE-MCL algorithm

TRIBE-MCL is an iterative algorithm, which operates on a directional graph. Each of the N nodes of the graph represents a protein sequence from the set we wish to cluster, while each edge length S_{ij}, i, $j=1,\ldots,N$, shows the similarity between protein sequences with index i and j, respectively. In other words, edge length S_{ij} indicates the likelihood of an evolutionary transition from protein i to protein j. The edge lengths are stored in the so called similarity matrix S, which obviously is a square shaped $N \times N$ matrix. Theoretically, the initial edge lengths can be computed using any sequence alignment method. However, the computation load of the algorithm will depend on the initial similarities, so a careful choice of the distance metrics is recommendable.

Generally, this initial similarity matrix does not have to be symmetrical. This may be treated as a problem or not. If one only wishes to cluster the sequences, that is, to find certain groups of proteins, which show high similarity within the cluster and low similarity between proteins of different clusters, then using a symmetrical similarity matrix is recommendable. Asymmetrical metrics and the similarity values it yields, however, may reveal the direction of evolution among the proteins situated within a given cluster. Consequently, making the similarity matrix symmetrical in every iteration is an extra step, whose benefits and costs have been tested in [15].

The TRIBE-MCL algorithm treats the similarity matrix as a stochastic matrix [8]. In such a matrix, estimated probability or possibility values are stored: in general, S_{ij} represents the possibility, that protein i becomes protein j during an evolutionary step. A decision has to be made at the beginning, whether S is treated as a column stochastic matrix or a row stochastic matrix. The difference is significant only from the biological point of view: if S is a column stochastic matrix, the columns are normalized in every iteration and thus they become estimated probability values of past evolutionary steps: for example S_{ij} shows what is the probability that protein j was i before the latest evolutionary step (e.g. mutation, insertion, etc.). On the other hand, rows of a column stochastic matrix are not normalized, so values in row j show the possible outcomes of next evolutionary steps, and their likelihood values, which are not probabilities as their sum is not 1. In this paper we chose to treat the similarity matrix as a column stochastic matrix.

The TRIBE-MCL algorithm consists of two main operations, namely the inflation and expansion, which are repeated alternately until a convergence is reached, that is, the similarity matrix becomes invariant during a cycle. Clusters are defined as connected subgraphs within the graph described by the similarity matrix, so a stable state of the similarity matrix means that the clusters do not change their contents during an iteration. Inflation has the main goal to differentiate among connections within the graph, by favoring more likely direct walks along the graph in the detriment of less likely walks. The strength of this differentiation is controlled by the so called inflation rate r: large values express the preference of likely walks more severely, causing sudden ruptures within the graph, possibly not in the ideal place. Low inflation rates are more likely to yield smooth partitions, but the convergence may become rather slow.

The other operation repeated in every iteration cycle is the expansion, which reveals possible longer walks along the graph. This is obtained via matrix multiplication, by computing the second power of the similarity matrix. This is a traditional way of finding longer walks within directional graphs. Here it is used to emphasize the likelihood of changes within the protein structures that happened in two or more evolutionary steps. This operation does not have a mean to control its strength, as we can only compute integer powers of the similarity matrix.

A simplified functional diagram of the original TRIBE-MCL algorithm is given in Fig. 1. Protein sequences taken from the SCOP95 database are aligned all against all using BLAST [1]. Similarity values are normalized and organized into a similarity matrix S, which is then fed to the clustering algorithm's main loop. After a certain number of iterations performed, when there is no relevant change detected in the similarity matrix during the latest iteration, the graph that corresponds to the final similarity matrix is interpreted, and clusters are established as connected subgraphs.

The original version of the TRIBE-MCL algorithm was found accurate and efficient in the identification of protein families from large databases [9]. In the followings, we will present some aspects of its imperfection, and will attempt to give solutions that

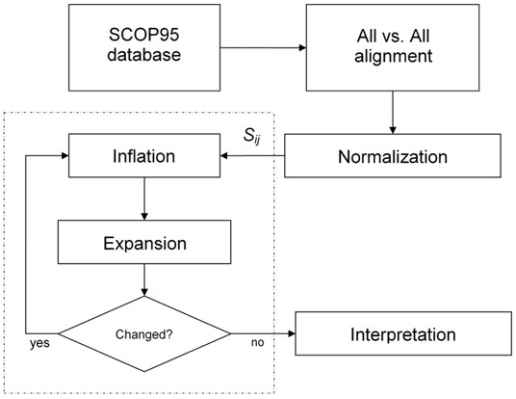

Fig. 1. Flowchart of the TRIBE-MCL algorithm. The main loop is indicated by the rectangle with dotted boundary.